Appendix B.
Sampling and Estimation Methodologies

The estimates in this report are based on a stratified simple random sample. The ICTS sample consists of 46,483 companies with paid employees (determined by the presence of payroll) in 2007.

The scope of the survey was defined to include all private, nonfarm, domestic companies. Major exclusions from the frame were government-owned operations (including the U.S. Postal Service), foreign-owned operations of domestic companies, establishments located in U.S. Territories, establishments engaged in agricultural production (not agricultural services), and private households.

The 2007 Business Register (BR) was used to develop the 2008 ICTS sample frame. The BR is the U.S. Census Bureau's establishment-based database. The database contains records for each physical business entity with payroll located in the United States, including company ownership information and current-year administrative data. In creating the ICTS frame, establishment data in the BR file were consolidated to create company-level records. Employment and payroll information was maintained for each six-digit North American Industry Classification System[1] (NAICS) industry in which the company had activity. Next, payroll data for each company-level record were run through an algorithm to assign the company, first to an industry sector (i.e., manufacturing, construction, etc.), then to a subsector (three-digit NAICS code), then to an industry group (four-digit NAICS code), then to an industry (five-digit NAICS code), and finally to an ICTS industry code based on the industry. The resulting sample frame contained slightly less than 6.0 million companies.

The 2008 ICTS sampling frame consists of a certainty portion and a noncertainty portion. The 17,964 companies with 500 or more employees were selected with certainty, meaning they are automatically part of the sample. The remaining companies with 1 to 499 employees were then grouped into 135 industry categories, and each industry further divided into four strata. Each stratum would be sampled from to build the rest of the sample. Since noncapitalized expenditures data were not available on the sampling frame, 2007 payroll was used as the stratification variable. The stratification methodology resulted in minimizing the sample size subject to a desired level of reliability for each industry. The expected relative standard errors (RSEs) ranged from 1 to 3 percent.

[1]Information about NAICS can be found at http://www.census.gov/eos/www/naics/

ESTIMATION

The quality measure called the sampling unit response rate is the percentage of all mailed eligible companies that responded, which is 79.9%. All companies are not equally important to the estimate however. Each sampled company has a sample weight reflecting other unselected companies in the population. Sampled companies in the same substratum have identical weights, which range from one, only represents itself, to several thousand. Respondents weights are further increased to widen their representation to account for companies that would have been

represented by nonrespondents. Final estimates use these increased weights. The coverage rate is a quality measure that is the percentage of the estimate of total noncapitalized and capitalized expenditures from respondents using only their original sampling weight. The coverage rate for ICTS was 91.7%. The difference between the two quality measures is that while many companies did not report to the survey, they are not as influential in creating the estimate as many who did report.

Sampling Weights and Weight Adjustment for Nonresponse

After being given an initial sampling weight, the weights could be further adjusted based on activity and response status. The goal is to have the in-scope responding sample reflect the entire in-scope frame. Each sampled company becomes a respondent, a nonrespondent, or out-of-scope due to being not in business during the survey year or a known duplicate to another record. Companies that went out of business during the survey year are still in-scope, and efforts are made to collect data for the period the company was active.

A company was considered a respondent or nonrespondent based on whether the company provided sufficient data in items 2 or 3 of the survey form asking about expenditures. Respondents will have their sampling weights adjusted upwards to account for nonrespondents, such that the weighted respondents still represent the entire in-scope population's total activity. The adjustment for respondents is based on the outstanding payroll that nonrespondents account for in each stratum. This adjustment may bias the estimates, since it is assumed that nonresponse is a purely random event, in that the relationship of payroll to expenditures does not differ in the aggregate from respondents to nonrespondents. No attempt was made to estimate the magnitude of any such bias due to the nonresponse weight adjustment.

In addition, companies who are deemed 'extreme outliers' may have their weights reduced to minimize their impact on the estimates and the mean squared error of the estimates.

ICTS segment. The following discussion assumes 675 strata (strata designation h = 1, 2, . . ., 675) which are based on 135 industries, each normally containing five strata (including the certainty stratum), which would be a maximum of 675 strata. When there is insufficient sample size to justify distinct strata, they are collapsed together.

The original stratum weights (W_h) were adjusted to compensate for nonresponse. The adjusted weight is computed as follows:

$$W_{h(adj)} = W_h * \frac{(P_{hr} + P_{hn})}{(P_{hr})}$$

where,

$W_{h(adj)}$is the adjusted stratum weight of the h^{th} stratum

$W_h = \frac{N_h}{n_h}$ is the original stratum weight of the h^{th} stratum

N_h is the population size of the h^{th} stratum
n_h is the sample size of the h^{th} stratum
P_{hr} is the sum of total company payroll for respondent companies in stratum h
P_{hn} is the sum of total company payroll for nonrespondent companies in stratum h

Cell Estimation

Publication cell or point estimates were computed by obtaining a weighted sum of reported values for companies treated as respondents.

ICTS segment. The ICTS estimates were derived as follows. Each estimated cell total, $\hat{X}_j$, is of the form

$$\hat{\chi}_j = \sum_{k=1}^{675} \sum_{i \in k} (W_{k(adj)} * \chi_{(j)i,k})$$

where,

$Wh_{(adj)}$ is the adjusted weight of the h^{th} stratum
$X_{(j)i,h}$ is the value attributed to the i^{th} company of stratum h,
where j is the publication cell of interest.

Note: Although a company was assigned to and sampled in one ICTS industry, it could report expenditures in multiple ICTS industries. When this occurred, the reported data for all industries were inflated by the weight in the sample industry.

RELIABILITY OF THE ESTIMATES

The values shown in this report are estimates from a sample and will differ from the data which would have been obtained from a different sample or a complete census. Two types of possible errors are associated with estimates based on data from sample surveys: sampling errors and nonsampling errors. The accuracy of a survey result depends not only on the measurable sampling errors but also on the nonsampling errors that are not explicitly measured. For any particular estimate, the total error may considerably exceed the measured sampling error.

Sampling Variability

The sample used in this survey is one of many possible samples that could have been selected using the sampling methodology described earlier. Each of these possible samples would likely yield different results. The relative standard error (RSE) is a measure of the variability among the estimates from all possible samples using this methodology. The RSEs were calculated using a delete-a-group jackknife replicate variance estimator. The RSE accounts only for sampling variability, and does not account for any nonsampling error or systematic biases in the estimates.

A bias is the difference, averaged over all possible samples of the same design and size, between the estimate and the true value being estimated.

The RSEs presented in the tables can be used to derive the standard error (SE) of the estimate. The SE can be used to derive interval estimates with prescribed levels of confidence that the interval includes the average results of all samples:

a. intervals defined by one SE above and below the sample estimate will contain the true value about 68 percent of the time.

b. intervals defined by 1.6 standard errors above and below the sample estimate will contain the true value about 90 percent of the time.

c. intervals defined by 2 standard errors above and below the sample estimate will contain the true value about 95 percent of the time.

The SE of the estimate can be calculated by multiplying the RSE presented in the tables by the corresponding estimate. Note, the RSE is the measure of variability presented for all estimates in this publication except for the estimates of percent changes presented in Table 2a[xls, 23KB], for which we provide the SE as the measure of variability (refer to Table 2b[xls, 22KB]). Also note that RSEs in this publication are in percentage form. They must be divided by 100 before being multiplied by the corresponding estimate.

Examples of Calculating a Confidence Interval:

a. For a data value: using data from Table 3a[xls, 25KB] and Table 3b[xls, 24KB], the SE for 2008 total nondurable manufacturing noncapitalized expenditures would be calculated as follows:

$$\hat{\sigma}(\hat{x}_j) = \left[\frac{RSE(\hat{x}_j)}{100}\right] * \hat{x}_j = \left(\frac{1.4}{100}\right) * \$5{,}390$$ million = $75.5 million

The 90-percent confidence interval can be constructed by multiplying 1.6 by the SE, adding this value to the estimate to create the upper bound, and subtracting it from the estimate to create the lower bound.

$$\hat{x}_j \pm \left[1.6 * \hat{\sigma}(\hat{x}_j)\right]$$

Using data from Table 3a[xls, 25KB], for 2008 total nondurable manufacturing noncapitalized equipment expenditures, a 90-percent confidence interval would be calculated as:

$5390 million ± 1.6*($75.5 million) = $5390 ± $121 million

This implies if there were repeated samples taken of the population, and similar confidence intervals created each time, 90 percent of those confidence intervals would contain the true population value, and 10 percent would not, if only uncertainty due to sampling is taken into

account. This is generally translated to having 90 percent confidence that the interval of $5,269 million to $5,511 million also contains the actual total for Nondurable Manufacturing noncapitalized equipment expenditures, subject to further nonsampling errors.

b. For percent change: using data from Table 2a[xls, 23KB] and Table 2b[xls, 22KB], the 90-percent confidence interval for percent change can be constructed by multiplying 1.6 by the SE of the percent change, adding this value to the estimated percent change to create the upper bound, and subtracting it from the estimate to create the lower bound. For example, for the noncapitalized expenditures in the Health care and social assistance sector, the estimated percent change from 2007 to 2008 is -6.9 percent (from Table 2a[xls, 23KB]), and the standard error of this estimate is 10.3 percent (from Table 2b[xls, 22KB]).

$$-6.9 percent \pm [1.6 * 10.3 percent] = -6.9 \pm 16.48 percent$$

The 90 percent confidence interval is then -23.38 percent to 9.58 percent change in this sector for noncapitalized expenditures. Since this interval contains zero (0), there is not sufficient evidence to conclude that the estimated percent change was statistically different from no change despite the point estimate being a negative number, i.e., the percent change is not statistically significant

Examples of Calculating Absolute Differences and Percent Changes

Data for the current year along with revised data for the prior year are presented in this publication. Two numbers of interest for many data users may be the absolute difference between the prior year and the current year, and the percent change from the prior year to the current year.

The absolute difference is calculated as:

$$\hat{d}_j = [\hat{x}_t - \hat{x}_{t-1}]$$

and a 90-percent confidence interval on this difference is estimated as:

$$CI(\hat{d}_j) = \hat{d}_j \pm 1.6 * \sqrt{[\sigma^2(\hat{x}_t) + \sigma^2(\hat{x}_{(t-1)})]}$$

As an example, for the capitalized equipment expenditures for computer and peripheral equipment in the Retail trade sector, from Table 3a[xls, 23KB], the estimate for 2008 is $$4,164 with the RSE found in Table 3b[xls, 25KB] as 2.7, and for 2007 the revised estimate from Table 3a [xls, 25KB] is $4,462 with the RSE found in Table 3b[xls, 25KB] as 2.8. The above calculations would be:

$$\hat{d}_j = [\$4{,}164 million - \$4{,}462 million] = \$298 million$$

And the 90-percent confidence interval is estimated as:

$$CI(\hat{d}_j) = \$298 \pm 1.6 * \sqrt{[(.027 * \$4{,}164)^2 + (.028 * \$4{,}462)^2]}$$

= $298 ±1.6*$168.07

= $298 ± 268.92

~ ($29 ,$567) million

The 90-percent confidence interval is $298 +/- $269 million, or $29million to $567 million.

The percent change is calculated as 100 multiplied by the ratio of the difference divided by the prior estimate.

So continuing with the example from above,

$$PC_j = 100 * \left[\frac{\hat{d}_j}{\hat{x}_{(t-1)}} \right]$$

$$= 100 * \left[\frac{\$298}{\$4{,}462} \right]$$

$$= 6.68 \, percent$$

The 90-percent confidence interval on this percent change is estimated as:

$$CI(PC_j) = PC_j \pm 1.6 * 100 * \left(\frac{\hat{X}_t}{\hat{X}_{t-1}} \right) * \sqrt{\left[\left(\frac{RSE(\hat{X}_t)}{100} \right)^2 + \left(\frac{RSE(\hat{X}_{t-1})}{100} \right) \right]^2}$$

$$= PC_j \pm 1.6 * 100 * \left(\frac{\$4{,}164}{\$4{,}462} \right) * \sqrt{[(.027)^2 + (.028)^2]}$$

$$= PC_j \pm 1.6 * 100 * 0.001412$$

$$= PC_j \pm 0.2259 \, percent$$

$$= 6.68 percent \pm 0.23 \, percent$$

so the 90-percent confidence interval is 6.68 percent +/- 0.23 percent or 6.45 percent to 6.91 percent.

Nonsampling Error

All surveys and censuses are subject to nonsampling errors, which are uncertainties in the estimates due to reasons other than taking a sample of the population for measurement. Nonsampling errors can be attributed to many sources: inability to obtain information about all

companies in the sample; inability or unwillingness on the part of respondents to provide correct information; response errors; definition difficulties; differences in the interpretation of questions; mistakes in recording or coding the data; and other errors of collection, response, coverage, and estimation for nonresponse. In addition, the sampling frame may have inaccuracies such as not including in-scope cases. These coverage errors may have a significant effect on the accuracy of estimates for this survey. The businesses that are on the sampling frame may also have outdated or inaccurate data, such as inaccurate payroll, which can influence the estimates.

Explicit measures of the effects of these nonsampling errors are not available. However, to minimize nonsampling error, all reports were reviewed for reasonableness and consistency, and every effort was made to achieve accurate response from all survey participants.